Title
Variable Selection using Stepwise Regression with Application to MyTherapistMatch.com

Author
Macchia, Francesco

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Variable Selection using Stepwise Regression with Application to MyTherapistMatch.com

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Francesco Macchia

2012

Abstract of the Thesis

Variable Selection using Stepwise Regression with Application to MyTherapistMatch.com

by

Francesco Macchia

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Frederic Paik Schoenberg, Chair

MyTherapistMatch.com is a website that connects patients with therapists who are deemed to be well suited to patients' personality and mental health needs. The primary tool for this patient-therapist matching is an online questionnaire that is completed by users visiting the site. Only approximately ten percent of visitors to the site complete the questionnaire, ostensibly partly due to the excessive length of the survey. The purpose of this thesis is to identify the items on the questionnaire that have the greatest impact on how much users interact with the website and reach out to their therapist matches. This is done with linear regression techniques, including ordinary least squares regression and stepwise regression. According to the regression analyses, the single greatest determinant of user activity is whether the respondent lives in or outside of California. The proprietors of the website may use the results of this analysis in choosing a meaningful subset of items that forms a shorter questionnaire.

The thesis of Francesco Macchia is approved.

Robert L. Gould

Nicolas Christou

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2012

To my mother, for believing in me;
to Cherine, for the years of friendship and encouragement;
and to Ramon, for your love and unwavering support

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the staff at MyTherapistMatch.com, especially Corey Quinn for allowing me the opportunity to work on this project, and Shannon Hughes for helping with the data files. I am deeply grateful to my advisor, Rick Schoenberg, for all the guidance and support he provided on this project. Thanks to Rob Gould and Nicolas Christou as well for being kind enough to be on my thesis committee. I would also like to thank Glenda Jones for having my back. And finally, my heartfelt gratitude to all my friends and family who have walked with me on this journey.

CHAPTER 1

Introduction

Visitors to the MyTherapistMatch.com website are asked to complete a questionnaire which is used to generate a personalized list of therapist matches. Users can then browse profiles of their therapist matches and retrieve contact information. A major problem that has been identified is that there is a high non-response rate on the part of users. That is, it appears that a high proportion of visitors to the website do not end up scheduling a session with a therapist. The major reasons for this appear to be:

- some users begin but do not complete the questionnaire, most likely due to its length
- some users receive no matches after completing the questionnaire
- some users are not satisfied with the quality of their matches

The purpose of this thesis is to determine how strongly each item on the questionnaire is associated with patients successfully finding therapists with whom they are compatible. Since the large number of items on the questionnaire seems to discourage some users from completing it, the results of our analysis may be used to reduce its length by eliminating items which have little effect on whether a patient ultimately finds a suitable match.

Follow-up surveys have been sent to users in an effort to gauge their satisfaction with their experience on the website. However, the response rate for this

survey has been extremely low. The number of completed follow-up surveys is currently far too small for these data to be used to evaluate the relative importance of questions. To this end, we used data on matches, clicks on therapists' contact information, and other recorded data indicating patients' utilization of the match information provided to them. We combined these into an overall measure of user activity, which serves as a proxy measure of user satisfaction.

CHAPTER 2

Methods

2.1 Data

The dataset for this project was provided by MyTherapistMatch.com. It consists of users' responses to the initial questionnaire, a log of user activity on the site once the questionnaire is completed, and users' responses to a follow-up survey regarding their experience on the website. All the data were collected from December 2009 through August 2011.

2.1.1 Initial questionnaire

The initial questionnaire was completed by 3,686 people. It consists of 58 items, 41 of which relate to various dimensions of personality. These dimensions, as identified by the website's proprietor, are "preferred representational system", "options/procedures", "towards/away", "internal/external", "proactive/reactive", "perceptual positions", "experience of time", "sameness/difference", and "specific/general". The following is an example from the "perceptual positions" group of items:

32. When expressing sympathy to someone who has lost a loved one, I feel:

    (a) my own sorrow.

    (b) the other person's sorrow.

    (c) that the other person's loss is unfortunate.

The remaining items are demographic questions (eg. date of birth, zip code, marital status) and questions regarding patients' therapy preferences (eg. therapists who offer online and/or tele-sessions). Each item on the questionnaire has a corresponding variable name beginning with the letter Q and followed by a one- or two-digit number (which does not match the item number). For example, item 32 above is called *Q56* according to the client's internal naming convention. All items on the questionnaire are listed with their corresponding variable names in the Appendix.

In order to facilitate my analysis, I created the following three new variables based on existing variables:

- Item 46 asks respondents to select the option that best describes their religion or spirituality. There are 38 response options to this item. For the sake of simplification, "Catholic peace traditions" was combined with "Catholic", "Orthodox Jewish" and "Reformed Jewish" were combined with "Jewish", and 23 Christian religions were combined with "Christian". Collapsing these response options resulted in a reduction from 38 to 12 levels for this item. This new variable is named *Rel*.

- Item 49 asks respondents for their birth date. Birth dates were used to calculate respondents' ages on August 1, 2001, regardless of the date the survey was taken. The new variable is named *Age*.

- Item 55 asks respondents for their zip code. Zip codes were used to designate respondents as either California or non-California residents. This new factor is named *State*. California residents account for 59.8% of all respondents.

2.1.2 User activity

MyTherapistMatch.com maintained and supplied a log of user activity on the site. For any given user, we can determine if and how many therapist matches were generated. We can also see if the user clicked on any of the links provided for each therapist. This includes the therapist's MyTherapistMatch.com profile, as well as links to the therapist's own website, email address, and phone number. A scoring system was devised to measure each user's activity on the site after completing the questionnaire, with varying numbers of points assigned for each type of action. The point values awarded for each action are listed in Table 2.1.

Action code	Action	Points
Match	A patient is matched with a therapist	1
MyVirtualShrink	A patient receiving no matches clicks on a link to another website specializing in computer-assisted therapy	3
ProfileView	A patient clicks on a link to a matched therapist's profile on MyTherapistMatch.com	3
WebsiteReferral	A patient clicks on a link to a matched therapist's own website	5
ContactClicked	A patient clicks on the "contact this therapist via email" link on a matched therapist's MyTherapistMatch.com profile	10
PhoneClicked	A patient clicks on the "contact this therapist via phone" link on a matched therapist's MyTherapistMatch.com profile	10

Table 2.1: Point values for user activity scoring system

In many cases, users performed the same action with the same therapist multiple

times. For example, a patient may have clicked on the same therapist's profiile several times. In some of my linear models, I assigned points for every instance of every action, while in others, I only assigned points for the first instance of a particular type of action with respect to a particular patient/therapist combination. Basic summary statistics for activity scores under both scenarios are listed in Table 2.2. Scores range from 0 to 632 for the scoring system with duplicate activity included and range from 0 to 181 for the scoring system with duplicate activity excluded. The former has a standard deviation of 30.66 while the latter has a standard deviation of 14.01. The contrast in the spreads of the two scoring systems is illustrated in the box plots of the two distributions in Figure 2.1.

Scoring method	Maximum score	Mean score	Median score	Standard deviation
Total points, all actions	632	20.0862	11	30.6562
Total points, no duplicates	181	12.1914	9	14.0078

Table 2.2: Summary statistics for activity scores

2.1.3 Follow-up survey

Two weeks after completing the initial questionnaire, all respondents are sent a follow-up survey. This survey was completed by only 72 people. The survey consists of six questions regarding users' impressions of their experience on the website. For the purposes of this analysis, we are concerned only with Question 2 of the follow-up survey:

2. How would you rate the quality of therapist matches you received on MyTherapistMatch.com?

   (a) Excellent – I found a great therapist

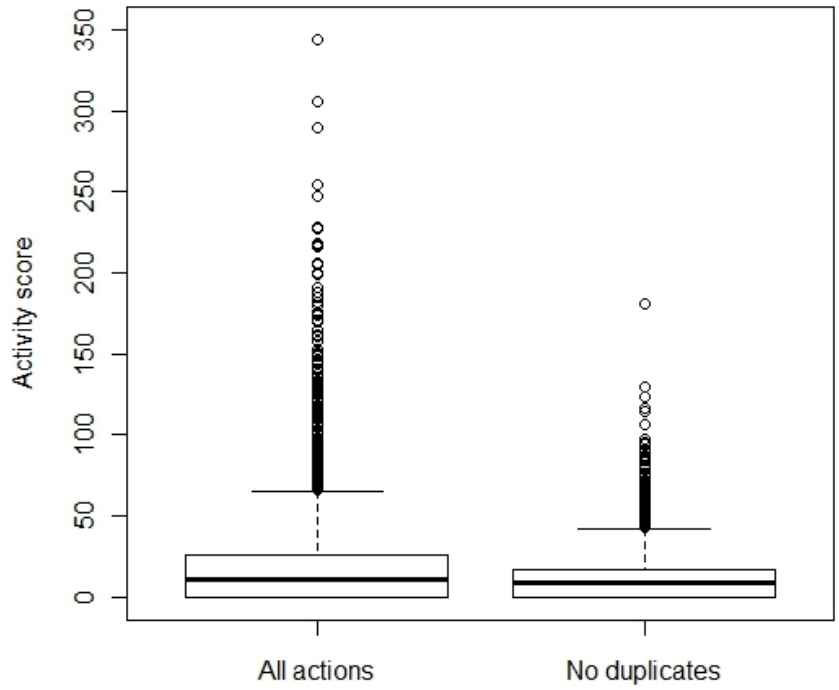

Figure 2.1: Median and interquartile ranges of activity scores

(b) Good – they seemed fine

(c) Needs help – none spoke to me

(d) Not sure yet

(e) Other (please specify)

For my analysis, I coded the response options "Excellent" and "Good" as positive and "Needs help" as negative. Where possible, users responding "Other" were coded as either positive or negative, depending on the comments they entered.

2.2 Regression analysis

Linear regression techniques were used to identify the questionnaire items having the greatest impact on user activity scores. These include ordinary least squares regression and stepwise regression.

2.2.1 Ordinary least squares regression

Generally, a linear model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

where y is the response variable, $x_1 \ldots x_n$ are the explanatory variables, β_0 is the intercept term, $\beta_1 \ldots \beta_n$ are the regression coefficients, and ϵ is a random error term. The ordinary least squares method of regression consists of choosing estimates of the β terms such that the sum of the squares of the errors is minimized.[1]

Most of the items on the initial questionnaire have multiple categorical response options. For each of these categorical items, the number of terms in the linear model will be one less than the number of response options for that item. The full linear model for this dataset can be expressed as

$$\begin{aligned} y \;=\; & \beta_0 + \beta_{Q3,10} x_{Q3,10} + \beta_{Q3,11} x_{Q3,11} + \beta_{Q3,12} x_{Q3,12} + \\ & \beta_{Q4,14} x_{Q4,14} + \beta_{Q4,15} x_{Q4,15} + \beta_{Q4,16} x_{Q4,16} + \ldots + \\ & \beta_{State,Non-CA} x_{State,Non-CA} \end{aligned}$$

Q3 is the first item on the questionnaire. The four response options for *Q3* are coded 9, 10, 11, and 12. In the linear model above, there are terms for all but the first response option. If a user chooses 9 as a response to *Q3*, all three regression coefficients related to *Q3* are set to zero. If a user chooses 10 as a response to *Q3*, the regression coefficient $\beta_{Q3,10}$ will be a non-zero value determined by the regression analysis, and the regression coefficients for response option 11 and 12 will be set to zero. The same principle is true if a user selects option 11 or 12, and also applies to every categorical variable on the questionnaire. Another example of this is the *State* variable, which is listed as the final variable in the linear model above. Respondents are classified as either California residents or non-California residents. Accordingly, there is just one term in the linear model for the *State*

variable. For California residents, the value of $\beta_{State,Non-CA}$ will be zero. For non-California residents, it will be a non-zero value determined by the regression analysis.

2.2.2 Stepwise regression

In ordinary least squares regression, all variables are evaluated at the same time. Stepwise regression is different in that variables are either included or excluded from the model one at a time. In backward stepwise regression, we start with the full model and eliminate the least significant variable. The model is re-fit to this subset of variables and the new least significant variable is eliminated. This procedure is applied iteratively until all non-significant variables have been removed from the dataset. In forward stepwise regression, we start with no variables in the model and add the most significant variable. From the remaining variables, the next most significant variable is selected and added to the model. This is repeated until a new variable does not sufficiently improve the fit of the model to justify its inclusion.[2]

CHAPTER 3

Results

3.1 Linear models

Several linear models were fit to various subsets of the questionnaire data. In some cases, I performed regression analyses on the dataset including all respondents, and in other cases, I performed regression analyses on the dataset including only California residents. For both groups of models, I used ordinary least squares regression and excluded the *State* variable from the analysis. In a third group of linear models, I included the *State* variable and included all respondents in the analyses. This third group of models included both ordinary least squares regression and stepwise regression. For each group of models, I attempted using both the full scoring system and the scoring system with no duplicate activity. I also introduced maximum activity scores as a way of mitigating the effects of outliers. These various combinations of data scenarios ultimately resulted in 21 different linear models.

Table 3.1 lists these models along with the following statistical information: residual standard error, adjusted R^2, p-value, and the number of variables that impact activity scores at the 5% level of significance. The residual standard error is a measure of the difference between the values predicted by the model and the actual observed data. A residual standard error of 12.48, for instance, indicates that for a typical respondent, the regression model will predict that user's activity to within approximately 12.48 points. The adjusted R^2 figure is a measure of how

well the model fits the data. Given multiple linear models, a lower residual standard error and a higher R^2 argue in favor of selecting a particular model.[3]

Data	Reg method	Activity score	Res std error	Adj R^2	P-value	# of variables
All clients	Ordinary least squares	All actions, no cap	30.30	0.0346	0.0000	11
		All actions, 100 pt cap	22.37	0.0567	0.0000	11
		No dup, no cap	13.67	0.0543	0.0000	9
		No dup, 100 pt cap	13.28	0.0573	0.0000	9
		No dup, 50 pt cap	11.60	0.0715	0.0000	10
CA clients only	Ordinary least squares	All actions, no cap	36.07	0.0062	0.1956	6
		All actions, 100 pt cap	25.06	0.0227	0.0024	8
		No dup, no cap	15.39	0.0155	0.0229	10
		No dup, 100 pt cap	14.82	0.0169	0.0158	9
		No dup, 50 pt cap	12.42	0.0262	0.0007	12
All clients with *State* included	Ordinary least squares	All actions, no cap	29.27	0.0989	0.0000	7
		All actions, 100 pt cap	21.23	0.1500	0.0000	10
		No dup, no cap	12.88	0.1575	0.0000	9
		No dup, 100 pt cap	12.48	0.1646	0.0000	10
		No dup, 50 pt cap	10.77	0.1949	0.0000	10
	Forward stepwise	No dup, no cap	12.88	0.1584	0.0000	12
		No dup, 100 pt cap	12.48	0.1659	0.0000	13
		No dup, 50 pt cap	10.77	0.1980	0.0000	15
	Backward stepwise	No dup, no cap	12.88	0.1575	0.0000	11
		No dup, 100 pt cap	12.48	0.1652	0.0000	12
		No dup, 50 pt cap	10.77	0.1972	0.0000	14

Table 3.1: Summary of the linear regression models.

The summary of the linear regression models indicates that both the omission of duplicate activity and the imposition of a cap on scores have the effect of reducing the residual standard error and increasing the adjusted R^2. The linear models for the dataset with all respondents and excluding *State* as a variable have residual standard errors ranging from 30.30 for the least constrained scores to 11.60 for the most constrained scores. The adjusted R^2 ranges from 0.0346 for the least constrained scores to 0.0715 for the most constrained scores. The linear models for the subset of data containing only California residents have residual standard errors ranging from 36.07 to 12.42 and adjusted R^2 ranging from 0.0062 to 0.0262. The linear models for the full dataset with *State* included as a variable have residual standard errors ranging from 29.27 to 10.77 and adjusted R^2 ranging from 0.0989 to 0.1980. While the minimum residual standard errors are comparable across the three groups of linear models, the full dataset with *State* as a variable nonetheless results in the lowest standard error. Larger differences are observed in the maximum adjusted R^2, and the full dataset with *State* as a variable results in the largest adjusted R^2. Due to these findings, I focus my analysis on the full dataset with *State* included as a variable, and on scores with duplicate activity omitted and constrained to a maximum of 50 points. The results that follow are based on the backward stepwise regression model for this dataset.

The backward stepwise regression analysis produces a model consisting of 14 variables. These variables are listed in Table 3.2 in order of significance. The question of whether a respondent lives in California is the single most significant predictor of one's activity score. Living in California has the effect, on average, of increasing a user's score, while the reverse is true for people living outside California. The next most significant variable is the question regarding interest in therapists who offer online or tele-sessions. This interest has the effect of increasing a user's score. The two most significant items with respect to personality are *Q17* and

Q20, both of which are of the Towards/Away type. The Towards/Away group of questions is the most heavily represented in this model, accouting for three of the eight significant personality items. The Sameness/Difference group of questions is represented by two items in this model.

The equation of the backward stepwise regression model can be expressed as

$$\begin{aligned} y \;=\; & \mathbf{17.1211} + 0.1619x_{Q4,14} + 1.0652x_{Q4,15} - 0.5679x_{Q4,16} \; \mathbf{- 3.1378x_{Q15,57}} \\ & + 0.2500x_{Q15,58} - 0.7997x_{Q15,59} \; \mathbf{- 1.2392x_{Q17,65}} + 0.2625x_{Q17,66} \\ & - 1.1633x_{Q17,67} - 0.7661x_{Q20,77} - 0.3217x_{Q20,78} \; \mathbf{+ 1.1702x_{Q20,79}} \\ & - 0.7382x_{Q36,117} \; \mathbf{+ 0.9250x_{Q38,121} - 0.7612x_{Q41,127} - 0.9983x_{Q76,217}} \\ & + 0.2046x_{Q76,218} + 0.3340x_{Q79,235} - 1.0758x_{Q79,236} + 3.3057x_{Q79,237} \\ & + 1.6073x_{Q79,238} \; \mathbf{+ 2.9402x_{Q79,239} - 1.6559x_{Q79,240}} - 1.6076x_{Q80,242} \\ & + 2.7215x_{Q80,243} + 0.6155x_{Q80,244} \; \mathbf{+ 3.3587x_{Q80,245}} + 0.3500x_{Q80,313} \\ & + 1.6777x_{Q80,314} - 1.5237x_{Q81,251} \; \mathbf{- 2.2562x_{Q81,253} - 5.0437x_{Q81,259}} \\ & - 0.3208x_{Q81,260} - 0.2570x_{Q81,263} - 1.2945x_{Q81,271} \; \mathbf{- 4.4411x_{Q81,274}} \\ & \mathbf{- 2.6257x_{Q81,275} - 2.5838x_{Q81,414} - 2.4115x_{Q81,415} - 0.8747x_{Q85,304}} \\ & - 0.8066x_{Q85,305} \; \mathbf{+ 5.0957x_{Q92,True} - 9.2425x_{State,Non-CA}} \end{aligned}$$

The value of the intercept term in this model is 17.1211. This would be the predicted activity score for a person completing the questionnaire who chooses the first response option for every item represented in the model, who declines interest in therapists who offer online and/or tele-sessions, and who resides in California. Responses deviating from this specific set of answer choices affect the predicted activity score according to their corresponding regression coefficients. Positive coefficients increase the predicted activity score and negative coefficients have the opposite effect.

The terms in the equation of the linear model in bold type are the specific answer

Variable	Item No.	Item Text	Item Type
State	55	My zip code (CA vs. non-CA)	(demographic)
Q92	57	Include therapists who offer online and/or tele-sessions	-
Q80	45	I am: (relationship status)	(demographic)
Q17	9	I seek personal relationships, in order to:	Towards/Away
Q20	12	What is likely to motivate you more?	Towards/Away
Q79	44	I am: (sexual orientation)	(demographic)
Q38	21	Regarding employment, I prefer to:	Sameness/Difference
Q76	40	I often think about activities I:	Experience of Time
Q81	46	I identify with the following religion/spirituality	(demographic)
Q41	23	When buying a car, I tend to prefer purchasing	Sameness/Difference
Q15	8	If I were to exercise, I would do so in order to:	Towards/Away
Q85	53	I drink alcohol?	(lifestyle)
Q36	20	If I were to buy a bird house that required assembly, I would:	Options/Procedures
Q4	2	I tend to communicate best with:	Preferred Representational Systems

Table 3.2: Items in the backward stepwise regression model, in order of significance.

choices that invidividually impact user activity scores at the 5% level of significance when all other variables are held constant. For the most part, it appears that the statistically significant answer choices are the ones whose regression coefficients have the largest absolute values. The *State* and $Q92$ variables, which are the most statistically significant in this model, not surprisingly have the largest regression coefficients in absolute value. However, this is not always necessarily the case. Using the variable $Q79$ as an example, the regression coefficient for the response option 237 is 3.3057 and the regression coefficient for response option 240 is -1.6559. The latter is statistically signficant, but the former is not. This is due to the number of respondents selecting each answer choice. Statistical signficance is determined by p-value, which is derived by calculating the t-statistic. The t-statistic is the coefficient divided by its standard error. A coefficient that is large relative to its standard error is an indication that a factor is more signficant than one for which the coefficient is not as large relative to its standard error. The calculation of standard error has as its denominator the number of observations. The larger the number of observations, the smaller the standard error, which makes the t-statistic larger. The smaller the number of observations, the larger the standard error, which makes the t-statistic smaller.[4] This explains the difference in statistical signficance of factors in spite of the absolute values of their regression coefficients. Going back to our example, $Q79$ is the item asking for respondents' sexual orientation. While 289 respondents selected option 240, "No comment", only 11 respondents selected option 237, "Transgendered". Even with a seemingly large coefficient, the number of users self-identifying as transgendered is too low, and thus the standard error too high, for this answer choice to be statistically signficant.

Figure 3.1 illustrates how the residual standard error changes as items are added to the selected model. The residual standard error of users' activity scores with

no explanatory variables is 12.02. The model consisting solely of the *State* variable and an intercept term improves the residual standard error by reducing it to 11.16. Adding *Q92* to the model results in further reduction of the residual standard error to 10.91. The addition of *Q80* to the model brings the residual standard error down to 10.88. This reduction continues until we reach the 14^{th} and final variable in the selected linear model. At this point, the residual standard error is 10.78, which is almost equal to the residual standard error of 10.77 of the full model containing all variables. This is in keeping with the notion that additional, non-statistically significant items do little, if anything, to improve the fit of the model to the data.

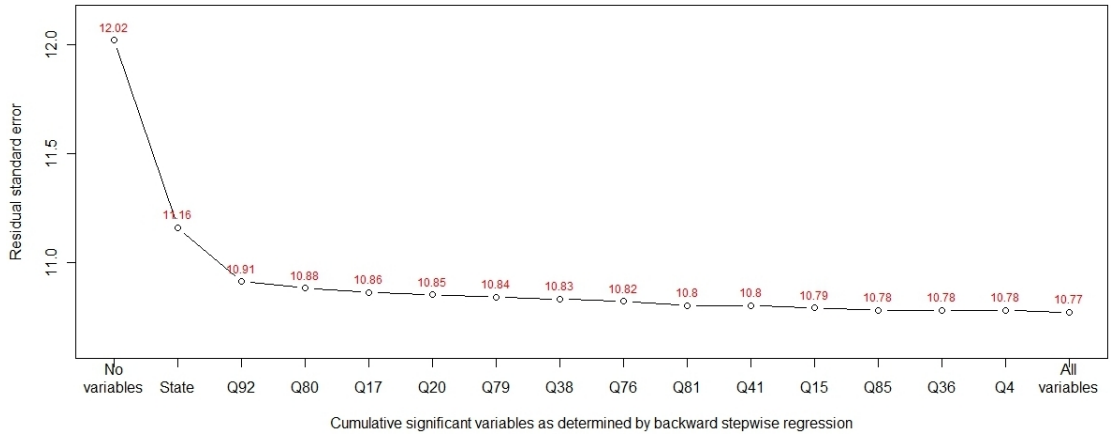

Figure 3.1: Residual standard error as variables are added to the backward stepwise regression model.

3.2 California versus non-California respondents

Due to the significance of the *State* variable in the chosen model, I compared summary statistics of the activity scores for California residents and non-California residents. Table 3.3 lists the maximum, mean, and median activity scores, as well as the standard deviation, for the two subsets of respondents under several

different scoring methods.

Data subset	Activity score	Maximum	Mean	Median	Standard deviation
CA clients only	All actions, no cap	632	27.3293	18	35.9701
	All actions, 100 pt cap	100	25.0971	18	25.3209
	No dup, no cap	181	16.2277	13	15.5074
	No dup, 100 pt cap	100	16.1501	13	14.9535
	No dup, 50 pt cap	50	15.5175	13	12.6400
Non-CA clients only	All actions, no cap	170	9.8220	4	16.0867
	All actions, 100 pt cap	100	9.6729	4	15.0165
	No dup, no cap	70	6.4717	3	8.8189
	No dup, 100 pt cap	70	6.4717	3	8.8189
	No dup, 50 pt cap	50	6.4422	3	8.6436

Table 3.3: Summary statistics of activity scores for California respondents and for non-California respondents.

There is a clear difference between the scores of California respondents and non-California respondents. In every scoring scenario, we see that California respondents scored higher than their out-of-state counterparts. For instance, when we look at scores with duplicate activity omitted, California residents have a mean score of 16.23 and a median score of 13, while non-California residents have a mean score of 6.47 and a median score of 3. When we further constrain the scores by imposing a cap of 50 points, Californians have a mean score of 15.52 and non-Californians have a mean score of 6.44. The difference in median values when we impose a 50 point maximum is illustrated in the box plots in Figure 3.2. They show the median and interquartile ranges of the activity scores for the two subsets of respondents. The thick horizontal bars represent the median score for each of

the two groups: 13 for California residents and 3 for non-California residents. The lower and upper borders of the rectangles represent the 25^{th} and 75^{th} percentiles, respectively. Graphically, the difference in the medians of the two groups is stark.

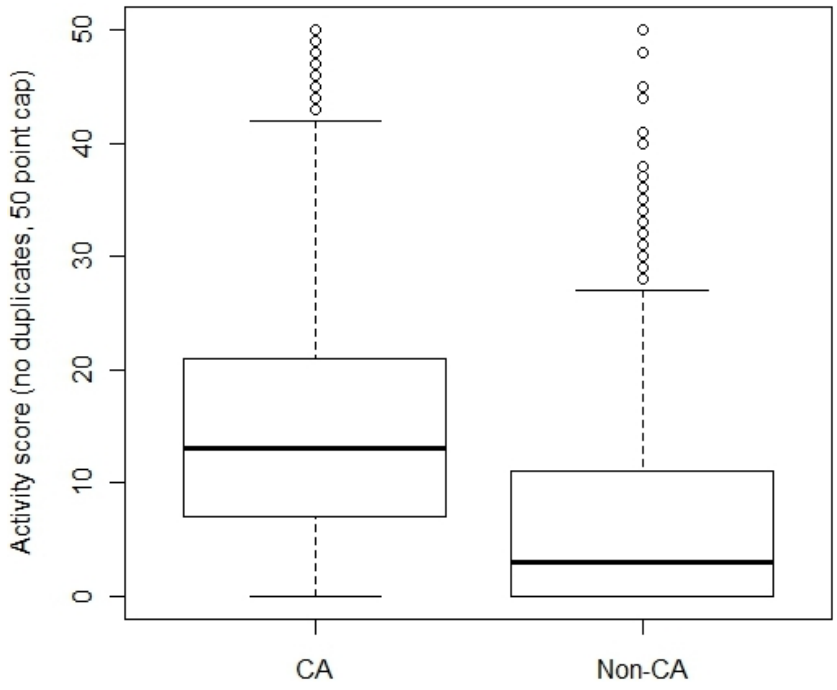

Figure 3.2: Median and interquartile ranges of activity scores of California respondents and of non-California respondents.

Another way to visualize the importance of the *State* variable is in the histograms in Figure 3.3. If we take 30 points as a cut-off, we can separate respondents into two groups: low scorers and high scorers. In the graph we can see that low scorers are split relatively closely between California and non-California residents. However, nearly 90% of the high scorers are California residents.

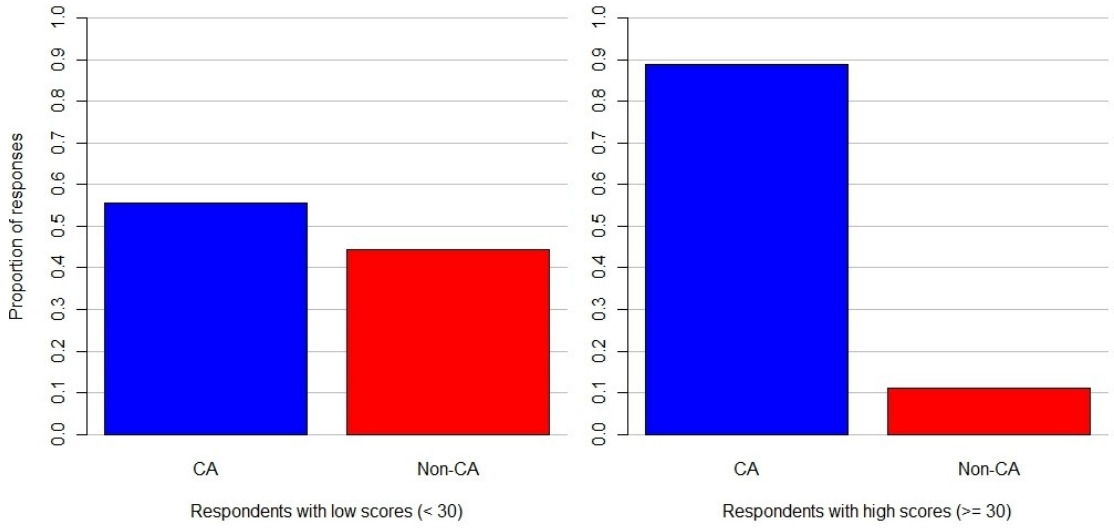

Figure 3.3: Proportion of respondents with low scores ($<$ 30) and high scores ($>=$ 30) who are California residents.

3.3 Follow-up survey

While the number of responses to the follow-up survey is low, we can nonetheless glean some general impressions from the data. The box plots in Figure 3.4 compare median activity scores for those responding negatively and those responding positively to Question 2 of the follow-up survey ("How would you rate the quality of therapist matches you received on MyTherapistMatch.com?"). The median action score is much larger for people who ultimately report being satisfied with their therapist matches. Patients responding positively to Question 2 on the follow-up survey had a median activity score of 15.5, while patients responding negatively had a median activity score of 9.5. This suggests that the scoring system employed in the regression analysis may be a useful proxy measure of user satisfaction.

Figure 3.5 plots the individual responses to Question 2 on the follow-up survey. While there is substantial scatter, it appears that positive resonses, plotted as filled circles, tend to be associated with higher scores relative to negative re-

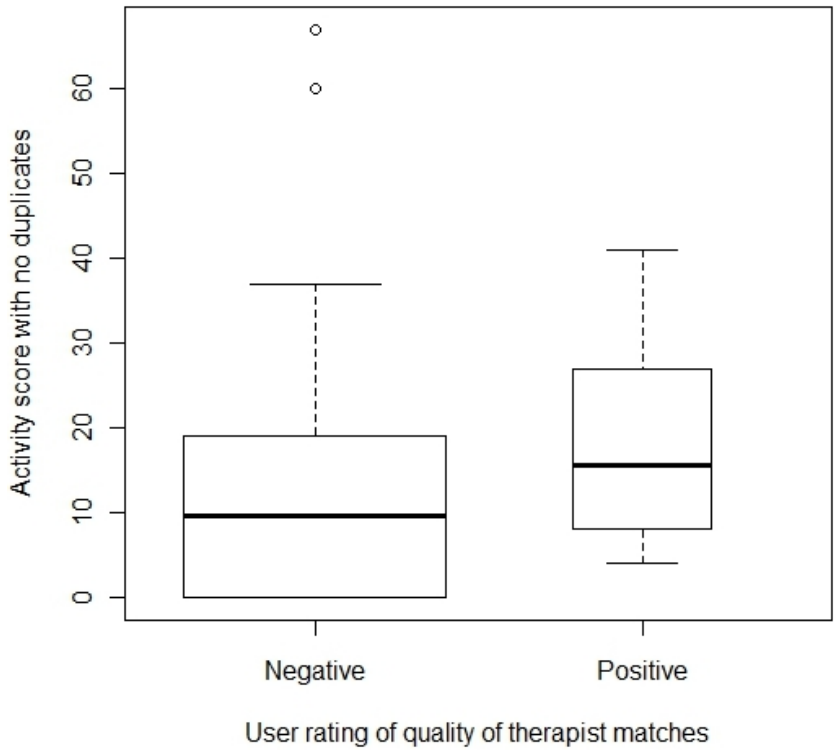

Figure 3.4: Median action scores by user rating of therapist matches.

sponses. The majority of the positive responses are from California residents (in blue) and the majority of the negative responses are from non-California residents (in red). This is in keeping with the finding that geographic location plays a large role in users' scores, and with the suggestion that it also plays a large role in user satisfaction.

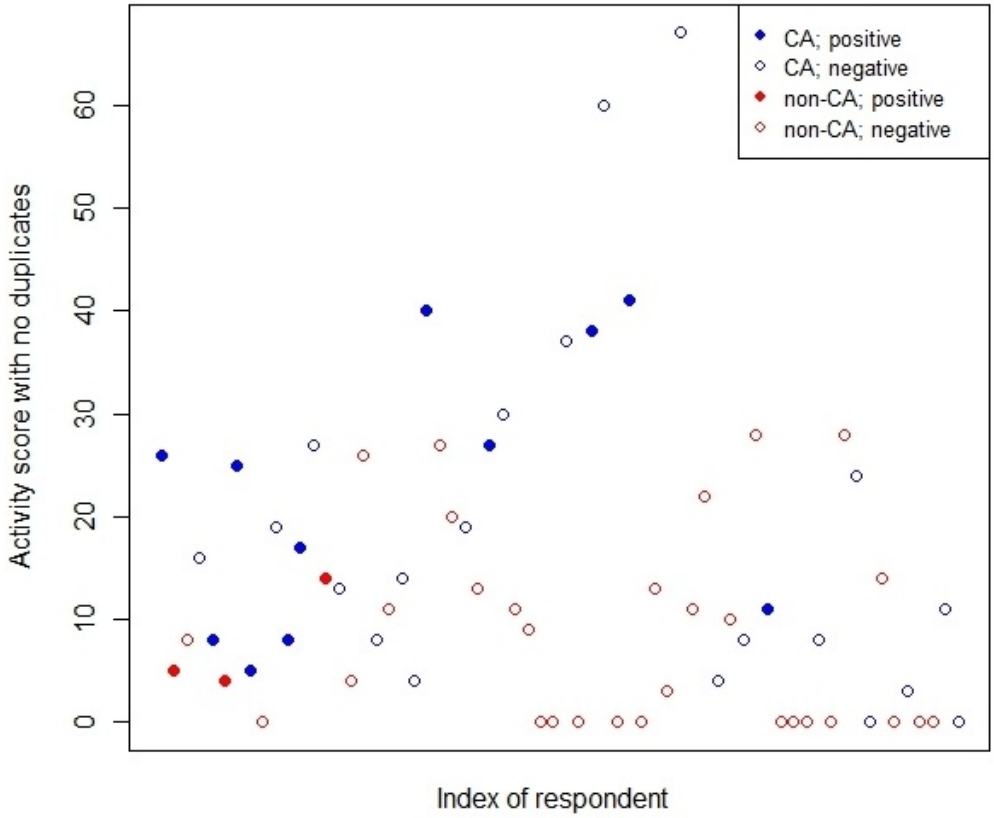

Figure 3.5: Scatterplot of activity score by respondents to the follow-up survey.

CHAPTER 4

Discussion

The goal of this project was to use statistical methods to generate a subset of items to be considered for inclusion in a revised, shorter questionnaire. My approach was to use linear regression techniques as a means of determining which questions most strongly correlate with user activity. The single most important factor affecting how much users followed up on their matches on the website was whether the user lived in California. This is apparently due in large part to the fact that the vast majority of therapists contracted with MyTherapistMatch.com at the time the data were collected are in California. A possible remedy to this problem is increasing the number of therapist listings generally and outside of California specifically. Another way to mitigate this problem is to alert visitors to the number of therapists in their area before they invest their time completing the questionnaire. The client has expressed an interest in implementing both of these plans.

The regression methods used in this project assume linearity. However, none of the linear models attempted here fit the data particularly well. In fact, the model providing the best fit to the data has an adjusted R^2 of just under 0.2, so other methods may be in order. In particular, future analysis might focus on exploring interaction effects between variables, as well as on the use of non-linear terms in the regression equation.

Another statistical technique that may help with identifying the most significant items on the questionnaire is cluster analysis. Cluster analysis attempts to summarize a data set meaningfully in terms of a relatively small number of groups of factors which resemble each other and which are different in some way from factors in other clusters.[5] Applied to this data set, cluster analysis could help identify redundancies in the questionnaire. If the response patterns to the questions in a cluster are similar to one another, the client could choose to keep only the most significant question from that cluster and discard the rest. Cluster analysis may also help to alert the client to any potential issues with construct validity. Construct validity, stated simply, is the principle that one is measuring that which one actually intends to measure.[6] The client has already conceived of groups of questions which are each meant to measure different dimensions of personality. We would expect clusters to roughly follow these theoretical partitions. For example, if a cluster analysis reveals an "experience of time" question amid a cluster of "internal/external" questions, this may indicate that responses to this particular "experience of time" item have more in common with responses to the "internal/external" items than with responses to other "experience of time" questions. The client could then explore the possibility that this item is not a valid or useful measure of the "experience of time" construct.

The ability to discern what is and what is not working with respect to the services provided by MyTherapistMatch.com is somewhat hampered by the sparsity of responses to the follow-up survey. While the user activity scoring system used here appears to be an effective proxy measure of patient satisfaction, direct feedback from users would likely provide a more accurate indication of patient satisfaction. The client is encouraged to find a way to increase the response rate to follow-up surveys. In the meantime, free response comments on the few completed follow-up surveys, such as the following from a patient in Long Beach, California, may be

instructive:

> *It's a great premise, and I liked the survey. I learned about myself taking it. I think you just need to get more therapists on board.*

Finally, it is crucial to note that while there is definitely a place for statistical analysis in the task of improving the initial questionnaire, it is not a substitute for the expertise of mental health professionals. The items whose answers optimally predict user activity are not necessarily the most important from a therapeutic perspective. Items identified by the regression analysis as being statistically signficant predictors of user activity deserve further consideration for inclusion on the questionnaire, but should not be understood to automatically form an appropriate collection of questions for matching patients with therapists.

CHAPTER 5

Appendix

5.1 Initial questionnaire items

Var. code	Item no.	Item text	Item type	Comments
Q3	1	When solving a problem, I tend to: (a) look at the big picture. (9) (b) consult with someone about it. (10) (c) get in touch with my deeper self. (11) (d) talk it over with myself or another person. (12)	Preferred Representational Systems	This item is a significant factor in 1 of the 15 OLS models.
Q4	2	I tend to communicate best with: (a) the volume and tone of my voice. (13) (b) logic. (14) (c) the way I look. (15) (d) my emotions. (16)	Preferred Representational Systems	This item is a significant factor in 2 of the 6 stepwise regression models.

Table 5.1: Initial questionnaire items.

Var. code	Item no.	Item text	Item type	Comments
Q7	3	I accomplish my work more easily if I: (a) clearly see what is wanted. (24) (b) have a feeling for what is required. (25) (c) talk with myself about what is needed. (26) (d) get instructions about what is wanted. (27)	Preferred Representational Systems	
Q8	4	It is easy to understand a presentation if: (a) I have hands-on experience. (28) (b) visual aids are used. (29) (c) it is based on logically presented ideas. (30) (d) the speaker emphasizes with tone and volume. (31)	Preferred Representational Systems	

Var. code	Item no.	Item text	Item type	Comments
Q9	5	I buy a car based on: (a) my thoughts about the price, mpg, and safety features. (32) (b) how it feels. (33) (c) its color, shape and look. (34) (d) the sound of the engine, the stereo system or how quiet it is. (35)	Preferred Representational Systems	This item is a significant factor in 1 of the 15 OLS models.
Q10	6	When talking with someone, I mostly notice: (a) whether or not the person is logical. (36) (b) the person's tone of voice. (37) (c) how I feel about the person. (38) (d) the person's body language and their point of view. (39)	Preferred Representational Systems	

Var. code	Item no.	Item text	Item type	Comments
Q12	7	I am good at: (a) seeing the big picture. (44) (b) understanding new facts and data. (45) (c) listening for what is right. (46) (d) embracing my feelings. (47)	Preferred Representational Systems	This item is a significant factor in 9 of the 15 OLS models.
Q15	8	If I were to exercise, I would do so in order to: (a) improve my health. (56) (b) avoid injury. (57) (c) get fit. (58) (d) avoid criticism from others. (59)	Towards/Away	This item is a significant factor in 3 of the 6 stepwise regression models.
Q17	9	I seek personal relationships in order to: (a) enjoy another's company. (64) (b) not be lonely. (65) (c) have my needs met. (66) (d) avoid isolation. (67)	Towards/Away	This item is a significant factor in 13 of the 15 OLS models and in 6 of the 6 stepwise regression models.

Var. code	Item no.	Item text	Item type	Comments
Q18	10	I brush my teeth to: (a) keep them healthy. (68) (b) avoid getting cavities. (69) (c) have a bright smile. (70) (d) reduce the chance of illness. (71)	Towards/Away	This item is a significant factor in 2 of the 15 OLS models.
Q19	11	When I wear my seatbelt, I do so to: (a) conform to the law. (72) (b) avoid a ticket. (73) (c) be safe. (74) (d) protect myself from injury. (75)	Towards/Away	
Q20	12	What is likely to motivate you? (a) working toward a goal. (76) (b) avoiding failure. (77) (c) achievement. (78) (d) fear of loss. (79)	Towards/Away	This item is a significant factor in 1 of the 15 OLS models and in 6 of the 6 stepwise regression models.
Q24	13	I know I've done a good job when: (a) someone lets me know. (92) (b) I notice it myself. (93)	Internal/ External	

Var. code	Item no.	Item text	Item type	Comments
Q26	14	When buying new clothes, I tend to buy whatever: (a) looks and/or feels right to me. (96) (b) my friends will probably like. (97)	Internal/ External	
Q30	15	I know I am right when: (a) I feel it in my gut. (104) (b) others tell me so. (105)	Internal/ External	
Q31	16	If I were to dance, I would do so: (a) to be seen. (106) (b) because it feels good. (107)	Internal/ External	This item is a significant factor in 4 of the 6 stepwise regression models.
Q32	17	When solving a problem, I prefer: (a) many alternatives. (108) (b) a step-by-step method. (109)	Options/ Procedures	
Q33	18	When cooking a meal, I tend to: (a) deviate from the recipe. (110) (b) follow the recipe. (111)	Options/ Procedures	

Var. code	Item no.	Item text	Item type	Comments
Q34	19	When planning a vacation, I prefer to: (a) create a detailed itinerary. (112) (b) figure out what to do when I arrive. (113)	Options/ Procedures	
Q36	20	If I were to buy a bird house that required assembly, I would: (a) follow the instructions. (116) (b) wing it. (117)	Options/ Procedures	This item is a significant factor in 6 of the 6 stepwise regression models.
Q38	21	Regarding employment, I prefer to: (a) be with the same employer for life. (120) (b) change employers or significantly change roles within the same company every two to three years. (121)	Sameness/ Difference	This item is a significant factor in 4 of the 15 OLS models and in 6 of the 6 stepwise regression models.

Var. code	Item no.	Item text	Item type	Comments
Q39	22	I prefer to live: (a) where I have roots. (122) (b) in various places, as it suits me. (123)	Sameness/ Difference	This item is a significant factor in 8 of the 15 OLS models.
Q41	23	When buying a car, I tend to prefer purchasing: (a) the same brand to stay with what works. (126) (b) a different brand to try something new. (127)	Sameness/ Difference	This item is a significant factor in 6 of the 15 OLS models and in 6 of the 6 stepwise regression models.
Q42	24	When going out to eat, I prefer eating at: (a) the same restaurant. (128) (b) new restaurants. (129)	Sameness/ Difference	

Var. code	Item no.	Item text	Item type	Comments
Q43	25	I agree with the following statement: (a) After attending a movie, I can tell a friend how the story unfolded. (130) (b) After attending a movie, I know if I liked it or not, but can't completely recall how the story unfolded. (131)	Specific/ General	
Q45	26	I agree with the following statement: (a) I generally prefer thinking about the big picture in life. (134) (b) I generally prefer thinking about particular details (people, places, things, etc.). (135)	Specific/ General	
Q46	27	At a restaurant, when paying the bill, I tend to: (a) review the bill closely, looking at all the details. (136) (b) just pay it. (137)	Specific/ General	

Var. code	Item no.	Item text	Item type	Comments
Q47	28	When involved in a misunderstanding, I tend to: (a) take initiative to solve the problem. (138) (b) wait for the other person(s) to approach me. (139)	Proactive/ Reactive	
Q49	29	When traveling with someone, I: (a) let others do the planning/ organizing. (142) (b) usually do the planning/organizing. (143)	Proactive/ Reactive	
Q52	30	When at work, I tend to: (a) be a self starter. (148) (b) wait for direction from others. (149)	Proactive/ Reactive	This item is a significant factor in 1 of the 15 OLS models.
Q53	31	When in an intimate relationship, I tend to: (a) be the first to express my feelings. (150) (b) let the other person express his/her feelings first. (151)	Proactive/ Reactive	

Var. code	Item no.	Item text	Item type	Comments
Q56	32	When expressing sympathy to someone who has lost a loved one, I feel: (a) my own sorrow. (156) (b) the other person's sorrow. (157) (c) that the other's loss is unfortunate. (158)	Perceptual Positions	
Q58	33	When I watch a sad movie, I: (a) feel sad about my life. (162) (b) feel sad for the characters in the movie. (163) (c) remind myself that it is just a movie. (164)	Perceptual Positions	
Q59	34	When I think of a painful event from my past, I: (a) relive my feelings as though it were happening now. (165) (b) think of the suffering the other person(s) went through. (166) (c) observe that event from a distance. (167)	Perceptual Positions	

Var. code	Item no.	Item text	Item type	Comments
Q61	35	When a friend gets injured, I: (a) think of my own pain. (171) (b) imagine his/her pain. (172) (c) mentally remove myself. (173)	Perceptual Positions	
Q63	36	When someone complains about a pain I've never experienced, I: (a) think I'm lucky that it didn't happen to me. (177) (b) try to imagine what he/she must be going through. (178) (c) think it's time for him/her to get over it. (179)	Perceptual Positions	
Q66	37	I often think about what: (a) I did in the past. (186) (b) I'm doing right now. (187) (c) I'll be doing in the future. (188)	Experience of Time	
Q73	38	I often think about people I: (a) used to know. (207) (b) currently know. (208) (c) want to know in the future. (209)	Experience of Time	

Var. code	Item no.	Item text	Item type	Comments
Q75	39	I often think about things I: (a) used to have. (213) (b) have now. (214) (c) want to have in the future. (215)	Experience of Time	
Q76	40	I often think about activities I: (a) used to engage in. (216) (b) do now. (217) (c) want to do in the future. (218)	Experience of Time	This item is a significant factor in 6 of the 6 stepwise regression models.
Q77	41	I often think about what I: (a) learned in the past. (219) (b) am learning now. (220) (c) will learn in the future. (221)	Experience of Time	

Var. code	Item no.	Item text	Item type	Comments
Q78	43	My ethnicity (select one) (a) White, non-Hispanic (222) (b) Hispanic or Latino (223) (c) African-American (224) (d) Asian / Pacific Islander (225) (e) Korean (226) (f) Japanese (227) (g) Chinese (228) (h) Indian (229) (i) Arab (230) (j) Native American (231) (k) Other (232) (l) No comment (233)	(demographic)	This item is a significant factor in 9 of the 15 OLS models.
Q79	44	I am (select one) (a) Heterosexual (234) (b) Homosexual (235) (c) Bisexual (236) (d) Transgendered (237) (e) Nonsexual (238) (f) Celibate (239) (g) No comment (240)	(demographic)	This item is a significant factor in 15 of the 15 OLS models and in 6 of the 6 stepwise regression models.

Var. code	Item no.	Item text	Item type	Comments
Q80	45	I am (select one) (a) Married (241) (b) Divorced (242) (c) Widowed (243) (d) Single (244) (e) No comment (245) (f) In a relationship (313) (g) Separated (314)	(demographic)	This item is a significant factor in 15 of the 15 OLS models and in 6 of the 6 stepwise regression models.
Rel (*Q81*)	46	I identify with the following religion(s)/spirituality (select one or more) (a) Buddhist (249) (b) Catholic (251) (c) Christian (253) (d) Hindu (259) (e) Islamic (260) (f) Jain (261) (g) Jewish (263) (h) Spiritual (271) (i) Not listed (274) (j) No comment (275) (k) Agnostic (414) (l) Atheist (415)	(demographic)	This item is a significant factor in 15 of the 15 OLS models and in 2 of the 6 stepwise regression models.

Var. code	Item no.	Item text	Item type	Comments
Q84	52	I smoke cigarettes? (select one) (a) Yes (300) (b) No (301) (c) No comment (302)	(lifestyle)	
Q85	53	I drink alcohol? (select one) (a) Yes (303) (b) No (304) (c) No comment (305)	(lifestyle)	This item is a significant factor in 15 of the 15 OLS models and in 4 of the 6 stepwise regression models.
Q86	54	I exercise (select one): (a) Rarely (306) (b) Sometime (307) (c) Frequently (308) (d) 7 days a week (309) (e) No comment (310)	(lifestyle)	This item is a significant factor in 3 of the 15 OLS models and in 2 of the 6 stepwise regression models.

Var. code	Item no.	Item text	Item type	Comments
Age (*Q87*)	49	My birth date (you must be 18 to use the features on this site)	(demographic)	This item is a significant factor in 4 of the 15 OLS models.
State (*Q90*)	55	My zip code	(demographic)	This item is a significant factor in 5 of the 5 OLS models in which *State* is a factor, and in 6 of the 6 stepwise regression models.
Q92	57	Include therapists who offer online and/or tele-sessions (a) False (b) True		This item is a significant factor in 14 of the 15 OLS models and in 6 of the 6 stepwise regression models.

References

[1] Douglas C. Montgomery. *Design and Analysis of Experiments,* 7^{th} *Edition.* John Wiley & Sons, Inc., 2009

[2] Julian J. Faraway. *Linear Models with R.* Chapman & Hall/CRC, 2005

[3] John Fox. *Applied Regression Analysis and Generalized Linear Models,* 2^{nd} *Edition.* Sage Publications, Inc., 2008

[4] Robert S. Witte, John S. Witte. *Statistics,* 9^{th} *Edition.* John Wiley & Sons, Inc., 2010

[5] Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. *Cluster Analysis,* 5^{th} *Edition.* John Wiley & Sons, Ltd., 2011

[6] Robert F. DeVellis. *Scale Development: Theory and Applications,* 3^{rd} *Edition.* Sage Publications, Inc., 2012